DOCUMENT RESUME

ED 310 136 TM 013 710

AUTHOR Rocklin, Thomas
TITLE Individual Differences in Item Selection in Computerized Self Adapted Testing.
PUB DATE Mar 89
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Attribution Theory; College Students; *Computer Assisted Testing; *Difficulty Level; Failure; Higher Education; *Individual Differences; Success; Test Anxiety; Test Items; *Test Wiseness
IDENTIFIERS *Preference Patterns; *Self Adapted Testing; Test Anxiety Inventory (Spielberger)

ABSTRACT

In self-adapted testing, examinees are allowed to choose the difficulty of each item to be presented immediately before attempting it. Previous research has demonstrated that self-adapted testing leads to better performance than do fixed-order tests and is preferred by examinees. The present study examined the strategies that 29 college students used in selecting items during a self-adapted test. After completing the Test Anxiety Inventory, subjects took the self-adapted test. The test contained 40 items sorted into 8 categories of difficulty based on Rasch model estimates. Three test-taking strategies were identified. Most subjects adopted a flexible strategy in which they generally selected easier items following failure and harder items following success. Some subjects adopted a "failure intolerant" strategy in which they generally selected easier items following failure and items of the same difficulty after success. Finally, some subjects adopted a "failure tolerant" strategy in which they chose items of the same difficulty level after failure, but harder items after success. The failure-tolerant strategy was associated with lower estimated ability than were the other two strategies. This finding may reflect the attributions examinees adopting that strategy make and the effort they expend following failures. The results provide general support for the value of continued development of self-adapted testing. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED310136

Individual Differences in Item Selection in Computerized Self Adapted Testing

Thomas Rocklin
374 Lindquist Center
The University of Iowa
Iowa City, IA 52242
319/335-5570
BITNET: CEDROCPA@UIAMVS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

THOMAS ROCKLIN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the 1989 annual meeting of the
American Educational Research Association, San Francisco, CA

TM013710

Abstract

In self adapted testing, examinees are allowed to choose the difficulty of each item to be presented immediately before attempting it. Previous research (Rocklin & O'Donnell, 1987) has demonstrated that self adapted testing leads to better performance than fixed order tests and is preferred by examinees. The present study examined the strategies that subjects used in selecting items during a self adapted tests. Three strategies were identified. Most subjects adopted a flexible strategy in which they generally selected easier items following failure and harder items following success. Some subjects adopted a "failure intolerant" strategy in which they generally selected easier items following failure and items of the same difficulty after success. Finally, some subjects adopted a "failure tolerant" strategy in which they chose items of the same difficulty after failure, but harder items after success. The failure tolerant strategy was associated with lower estimated ability than the other two strategies. This may reflect the attributions examinees adopting that strategy make and the effort they expend following failure.

BEST COPY AVAILABLE

Individual Differences in Item Selection in Computerized Self Adapted Testing

Thomas Rocklin

Test constructors, including classroom instructors, often begin a test with a few relatively easy items with the intent of minimizing the impact of examinees' test anxiety on their performance. Indeed, this practice is suggested in a number of text books on measurement (e.g., Kaplan & Sacuzzo, 1982). Specifying the difficulty of the first few items is a very simple approach to item sequencing[1] within a test. More complete specifications might include item sequences based on monotonically increasing difficulty, monotonically decreasing difficulty, "spiraling" difficulty, or random assignment of items to positions within the test. The effects of each of these item sequencing specifications on examinee performance has been investigated, but with mixed results (Lafitte, 1984).

Nonetheless, it seems likely that item sequencing makes a difference in examinee performance. When item difficulty is manipulated between tests instead of within tests, it interacts with examinee test anxiety in influencing performance on the examination (Rocklin & Thompson, 1985). In particular, the relationship between performance and test anxiety appears to be negative for difficult tests, but positive or curvilinear for easier tests. Given that item difficulty and examinee test anxiety appear to interact, the relation between item sequencing and performance is likely to be complex.

---

[1]Throughout this paper, "item sequencing" refers to sequencing in terms of item difficulty, rather than sequencing based on curriculum, objectives, or other content factors.

Testing technologies differ in the extent to which the test constructor has control over the sequence in which items are attempted. In traditional paper and pencil tests, the test constructor can select the order in which items are presented, and therefore exert modest control over the order in which they are attempted. Examinee's, however, normally have the ultimate control over the sequence in which items are attempted because they have the option to skip items that they find too difficult. Many books on "college survival" (e.g., Kesselman-Turkel & Peterson, 1981, American College Testing Program, 1989) contain advice to do just this. In any situation in which the examinee does not attempt all items on an examination, this strategy means that the examinee will actually adjust not only the item sequencing, but the overall test difficulty as well. Presumably, examinees make decisions about item sequencing based partly on their ability, and partly on current motivational and affective states. For example, an examinee who is feeling particularly anxious may seek a very easy item to gain confidence. On the other hand, a very calm examinee may enjoy the challenge of a difficult item.

Two kinds of test difficulty can be distinguished. The first is objective, or psychometric, difficulty. This is simply the average item difficulty, defined in terms of the proportion of examinees passing the item or in terms of an item response theory difficulty parameter. The second kind of difficulty, subjective difficulty, is likely to be more important to the examinee's motivation. Subjective difficulty is based on an examinee's perception of the probability that he or she has answered the item correctly.

Computerized adaptive testing (CAT) gives the examinee no control over the sequence in which items are attempted. In general, only one item is

available at a time and the difficulty of that item is selected algorithmically based on the examinee's previous responses. In CAT, although the objective difficulty of the examination will differ from examinee to examinee, the number of items answered correctly, and therefore the subjective difficulty, will be relatively constant, depending on the item selection algorithm and the item format (i.e., number of alternatives in a multiple choice item).

Thus, CAT provides tests that are individually tailored to examinees' ability levels, but are, in contrast to traditional paper and pencil tests, completely insensitive to individual examinees' motivational and affective states. In an attempt to allow examinees to make choices about item sequencing based on relatively full information, I have explored the potential of a technology I call self adapted testing (SAT; Rocklin & O'Donnell, 1987)). In SAT, examinees take a computer administered test, but instead of attempting items selected by algorithm, each examinee specifies the difficulty of the items he or she attempts on an item by item basis.

In making item selections, examinees taking a SAT have access to two kinds of information that is not generally available to examinees taking a paper and pencil test. First, the examinee receives item by item feedback, so that the subjective estimate of difficulty he or she makes is better informed. Second, the examinee is provided with normative information about the objective difficulty of the items from which he or she is to choose. Thus, in SAT, in contrast to paper and pencil testing, the examinee has access to the information necessary to make "rational" choices in item sequencing.

The success of SAT depends on examinees' ability to select item difficulties in ways that optimize their performance. In the initial evaluation of

SAT (Rocklin & O'Donnell, 1987), examinees randomly assigned to take a SAT performed better (i.e,, had higher ability estimates) than subjects taking either a relatively easy test or a relatively difficult test. In addition, there was no overall loss of precision of measurement associated with SAT, although there was an interaction between examinee's test anxiety and type of test.

Nearly all subjects in that study (86%) progressed from easier items at the beginning of the test to more difficult items at the end. Beyond this, though, little is known about the strategies examinees used for item selection. The purpose of this study was to examine these strategies in detail. In particular, the study was designed to answer these questions: (1) Can the item selection strategies of examinees be well-modeled using simple rules? (2) What are these rules? (3) Are the item selection strategies associated with the level or variability of examinees' performance or with examinees' test anxiety? In addition, SAT provides an environment for evaluating examinees' item sequencing preferences that might be relevant to other testing technologies.

Method

This study is based upon data collected in a previous study (Rocklin & O'Donnell, 1987). Subjects (university students) were recruited through campus wide advertisements offering $5.00 for participation in a one hour experiment and randomly assigned to take a hard, an easy, or a self adapted test based on the verbal section (analogies, antonymns, and synonymns in a five alternative multiple choice format) of the Scholastic Aptitude Test (College Entrance Examination Board, 1980). Only the data from the 29 subjects assigned to the self-adapted test are considered in this study.

After completing the Test Anxiety Inventory (Spielberger, 1980), subjects took the self adapted test. Forty items were sorted into eight categories of difficulty based on Rasch model estimates computed from previously collected data. Subjects specified the difficulty of the item they wished to attempt, responded to an item selected from that category, and were informed whether or not their response was correct. If no new items were available in the category, the subject was directed to choose another category. The test ended when 20 items had been answered or 10 minutes had elapsed, whichever came first.

Results

Three simple models of item selection strategies were evaluated. In each, item selection was guided by whether the previous items was answered correctly. No attempt was made to model selection of the first item. In the "failure tolerant" model, the examinee was assumed to choose an item of the same difficulty as the previous item following an incorrect response and an item of the next higher difficulty following a correct response. In the "failure intolerant" model, the examinee was assumed to choose an item of the next lower difficulty as the previous item following an incorrect response and an item of the same difficulty following a correct response. In the "flexible" model, the examinee was assumed to choose an item of the next lower difficulty as the previous item following an incorrect response and an item of the next higher difficulty following a correct response. Each model was used to simulate a response vector for each subject. This vector was the same length as the actual response vector and constrained by the availability of only 5 items in each difficulty category. The goodness of fit of each model was eval-

uated by computing the square root of the mean squared difference between corresponding elements in the two vectors.

For each subject, the best, second best, and worst fitting model were identified. For the best fitting model for each subject, goodness of fit ranged from .8 to 3.1 with a mean of 1.9 (on a scale of 1 to 8, corresponding to the 8 difficulty categories available to the subjects). Thus, some subjects behaved very much as one of the models predicted, while others were somewhat more idiosyncratic.

The mean goodness of fits and number of subjects best fit by each model are shown in Table 1. Most, but not all, subjects were best fit by the flexible model. Those subjects who were fit by the failure intolerant model were worse fit than those fit by other models ($F[2, 26]= 4.25$, $MS_e=.296$).

The ability of each subject was estimated from a one parameter model using item difficulties estimated from previously collected data(Wright, 1977). The mean ability estimates and mean standard errors of those estimates are shown in Table 1. The ability estimate means differ significantly ($F[2, 26] = 4.96$, $MS_e = 1.16$), with subjects best fit by the failure tolerant model receiving the lowest mean ability estimates. The mean standard errors do not differ significantly from one another.

Finally, the test anxiety scores (from the TAI), as shown in Table 1, do not differ significantly from one another.

Discussion

Given the information available to the examinee and the sole goal of estimating his or her ability, the most "rational" of the three strategies evaluate here is the flexible strategy. In fact, examinees adopting the flexible strat-

egy are essentially administering themselves a stradaptive test (Weiss, 1973). This strategy, in which incorrect answers are followed by attempts at easier items and correct answers are followed by attempts at harder items, was adopted by 18 (62%) of the subjects in this study. The other 11 subjects were better fit by a model in which difficulty was adjusted only after an incorrect answer (14% or 4 subjects) or only after a correct answer (24% or 7 subjects). These 11 subjects must have (a) had goals different from or in addition to the goal of estimating ability, (b) attended to information other than item difficulty (e g , their emotional state), or both (a) and (b).

The flexible strategy and the failure intolerant strategies were both associated with better performance than the failure tolerant strategy. Because there are so few examinees who selected items using the failure tolerant strategy, it is difficult to draw firm conclusions about them. They do not stand out in terms of test anxiety or any of the other attributes assessed in this study.

It seems plausible that these examinees are "failure avoiding" students (Covington & Omelich, 1985). Failure avoiding students (as opposed to success oriented and failure accepting students) have responded to repeated academic failure by trying to avoid responsibility for their failures. Thus, in this study, when they failed an item, they selected an equally hard item because they could then attribute their failure to the difficulty of the item, rather than their own low ability. When they answered an item correctly, they selected an item of the greater difficulty, to insure that if failure ensued it could be attributed to the item difficulty.

Although further research will be required to understand why examinee's make the item sequencing choices that they make, the results of this study provide strong evidence that there are indeed individual differences in item sequencing preferences. Given relatively full information, examinees differ in the seqeunce in which they want to attempt test items. Further, there appear to be at least two item sequencing strategies (the flexible and the failure intolerant) that are associated with equally good performance. Examinees who chose the failure intolerant strategy in this study would presumably find a typical CAT, which more closely resembles the flexible strategy, inhospitable. These examinees appreciate the chance to savor success. The present study does not establish a causal connection between item sequencing strategy and performance, but it seems likely that examinees who prefer the failure intolerant strategy would do more poorly in a CAT than a SAT.

The study reported in this paper provides general support for the value of continued development of SAT. In particular, the fact that not all examinees taking a SAT choose the same sort of item sequence, combined with the lack of evidence for superior performance being associated with any particular sequence, implies that examinees can take advantage of the self-tailoring afforded by SAT. It is also clear from questionnaire data (Rocklin & O'Donnell, 1987) that they appreciate the opportunity to make item sequencing choices.

What, finally, does the study reported in this paper tell us about the general issue of item sequencing? It is unlikely that there is a single item sequence for conventional tests that is optimal for all examinees. If the goal of the test constructor is to improve the performance of all examinees, he or she might be best off making the difficulties of items explicit examinees (e.g., by

grouping items into hard, medium and easy sections on the test form) and allowing examinees to make their own sequencing decisions.

Table 1

Characteristics of Subjects Best Fit by Each Model

	Failure Intolerant	Failure Tolerant	Flexible
Goodness of fit			
Mean	2.60	1.62	1.86
SD	.61	.50	.55
Estimated Ability			
Mean	1.00	-.58	.85
SD	.75	1.03	1.14
Std. Error of Ability Estimate			
Mean	.58	.60	.56
SD	.04	.08	.05
Test Anxiety			
Mean	39.25	39.86	35.83
SD	15.11	8.03	8.06
N	4	7	18

References

American College Testing Program (1989). *Building better study skills*. Iowa City, IA: ACT.

College Entrance Examination Board (1980). *An SAT: Test and technical data for the Scholastic Aptitude Test administered in March 1980*. Princeton: Educational Testing Service.

Covington, M. V., & Omelich, C. L. (1985). Ability and effort valuation among failure-avoiding and failure-accepting students. *Journal of Educational Psychology, 77*, 446-459.

Kaplan, R. M., & Sacuzzo, D. P. (1982). *Psychological testing: Principles, applications, and issues*. Monterey, CA: Brooks Cole.

Kesselman-Turkel, J., & Peterson, F. (1981). *Test-taking strategies*. Chicago: Contemporary books.

Laffitte, R. G. Jr (1984). Effects of item order on achievement test scores and students' perception of test difficulty. *Teaching of Psychology, 11*, 212-213.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315-319.

Rocklin, T., & Thompson, J. M. (1985). Interactive effects of test anxiety, test difficulty, and feedback. *Journal of Educational Psychology, 77*, 368-372.

Spielberger, C. D. (1980). *Preliminary professional manual for the Test Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Weiss, D. J., (1973) The stratified adaptive computerized ability test (Research Report 73-3). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Wright, B. D. (1977). Solving measurement problems withe Rasch model. *Journal of Educational Measurement, 14*, 97-116.